D428–D432 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki072

Reactome: a knowledgebase of biological pathways

G. Joshi-Tope[1,*], M. Gillespie[1,3], I. Vastrik[2], P. D'Eustachio[1,4], E. Schmidt[2], B. de Bono[2], B. Jassal[2], G.R. Gopinath[1], G.R. Wu[1], L. Matthews[1], S. Lewis[5], E. Birney[2] and L. Stein[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA, [2]European Bioinformatics Institute, Hinxton, Cambridge, UK, [3]St Johns University, NY, USA, [4]New York University School of Medicine, NY, USA and [5]University of California, Berkeley, CA, USA

Received August 19, 2004; Revised and Accepted October 6, 2004

ABSTRACT

Reactome, located at http://www.reactome.org is a curated, peer-reviewed resource of human biological processes. Given the genetic makeup of an organism, the complete set of possible reactions constitutes its reactome. The basic unit of the Reactome database is a reaction; reactions are then grouped into causal chains to form pathways. The Reactome data model allows us to represent many diverse processes in the human system, including the pathways of intermediary metabolism, regulatory pathways, and signal transduction, and high-level processes, such as the cell cycle. Reactome provides a qualitative framework, on which quantitative data can be superimposed. Tools have been developed to facilitate custom data entry and annotation by expert biologists, and to allow visualization and exploration of the finished dataset as an interactive process map. Although our primary curational domain is pathways from *Homo sapiens*, we regularly create electronic projections of human pathways onto other organisms via putative orthologs, thus making Reactome relevant to model organism research communities. The database is publicly available under open source terms, which allows both its content and its software infrastructure to be freely used and redistributed.

INTRODUCTION

Although sequencing of the human genome has proved to be a powerful tool for understanding biology, analysis of the genome has underscored the difficulty of deriving from it the higher principles of biology. Further, studying whole transcriptional profiles and cataloging protein–protein interactions has yielded much valuable biological information. Yet it remains difficult, often impossible, to make the leap from the genome or proteome to the physiology of an organism, an organ, a tissue or even a single cell.

The information that describes genes, their protein products, and the biological processes in which they are involved is scattered over several databases, the primary research literature and other publications. The inability to manipulate this knowledge computationally is most keenly felt in the analysis of high-throughput data. When a researcher looks at a set of microarray expression results, can he/she reliably notice that the dozens of up-regulated genes include all the components of the phosphodiesterase signal pathway? When a researcher identifies four quantitative trait loci for brittle bones in rats, will he/she realize that these four genomic regions all contain components of a named developmental pathway? The Reactome database was created to provide an integrated view of biological processes, which links such gene products and can be systematically mined by using bioinformatics applications. We curate well-established information about biological processes, and usually defer the curation of contentious data for curation to a later date, when clear supporting evidence may become available.

The primary curational domain for Reactome includes pathways from *Homo sapiens*, except when there are gaps in human data. Also, we regularly create electronic projections of human pathways onto other organisms via putative orthologs, thus making Reactome relevant to model organism research communities. Both Reactome content and software is publicly available under open source terms. Reactome supersedes an earlier project, the Genome Knowledgebase (1).

DESIGN AND IMPLEMENTATION

We first describe the Reactome user interface and then its underlying data model and data entry and curation processes.

User interface

The top of the Reactome homepage (Supplementary Material 1) shows the Reaction Map (Figure 1), a graphical summary of all

*To whom correspondence should be addressed. Tel: +1 516 367 6904; Fax: +1 516 367 6851; Email: joshi@cshl.org

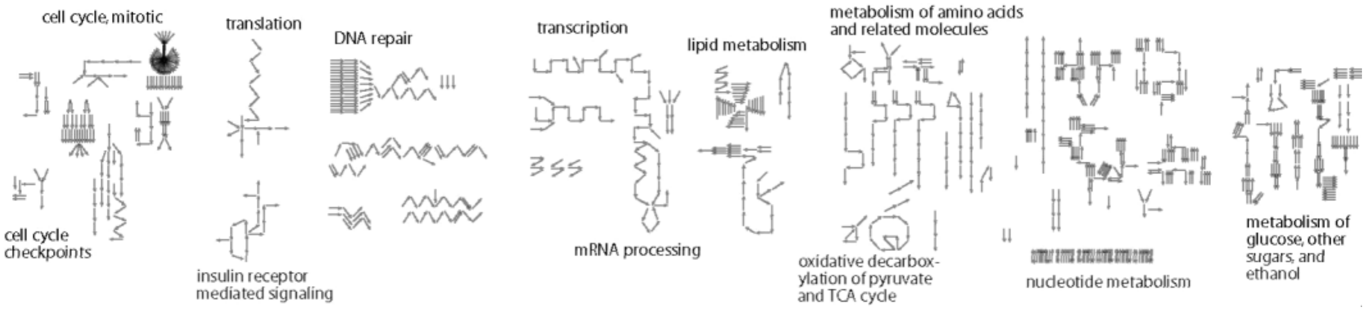

cell cycle, mitotic
translation
DNA repair
transcription
lipid metabolism
metabolism of amino acids and related molecules
cell cycle checkpoints
insulin receptor mediated signaling
mRNA processing
oxidative decarboxylation of pyruvate and TCA cycle
nucleotide metabolism
metabolism of glucose, other sugars, and ethanol

Figure 1. The Reaction map shows the reactions annotated in Reactome. The reaction clusters of top-level processes are shown.

the reactions contained within the database. Each reaction is represented as an arrow. Reactions that are causally or temporally related in a clustered set of reactions are arranged in a head-to-tail manner; thin gray lines indicate similar such links among reactions in different pathways. We have clustered the reactions to create recognizable patterns so that a researcher quickly learns to recognize each pathway's distinguishable appearance. This is useful both as a navigational tool to keep the biologist oriented as he/she moves around the resource, and also as a data-mining tool using the 'skypainter' facility, described in detail later.

As a researcher mouses over the reaction map, the relevant topic headers in the lower index panel are highlighted. The index panel lists a growing number of high-level processes among the six organisms for which events have been curated (human, Hsa) or inferred (mouse, Mmu; rat, Rno; puffer fish, Fru; zebra fish, Dre; and chicken, Gga). Each high-level process links to a chapter-like collection of pathways and reactions. A color-coding system allows the researcher to distinguish between reactions that have been curated from direct evidence in the literature from those that have been inferred via orthology in other species.

A researcher can navigate Reactome either by clicking on the reaction map, by selecting a topic of interest from a hierarchical table of contents, or via several structured database searches. As the researcher follows the path down into the resource he/she is presented with increasing levels of detail associated with the pathway: its constituent reactions, the participating complexes and macromolecules and the relationships among the pathways in the several species covered by Reactome. At each level, there is information designed for human browsing, such as text summaries of the pathway and hand-drawn diagrams of key events, as well as machine-readable information. As an example of the latter, Reactome makes it easy to obtain a list of all UniProt (2) accession numbers corresponding to proteins that participate in any phase of the cell cycle, or even in the cell cycle as a whole. A typical top-down browse path is shown in Supplementary Materials 2–5.

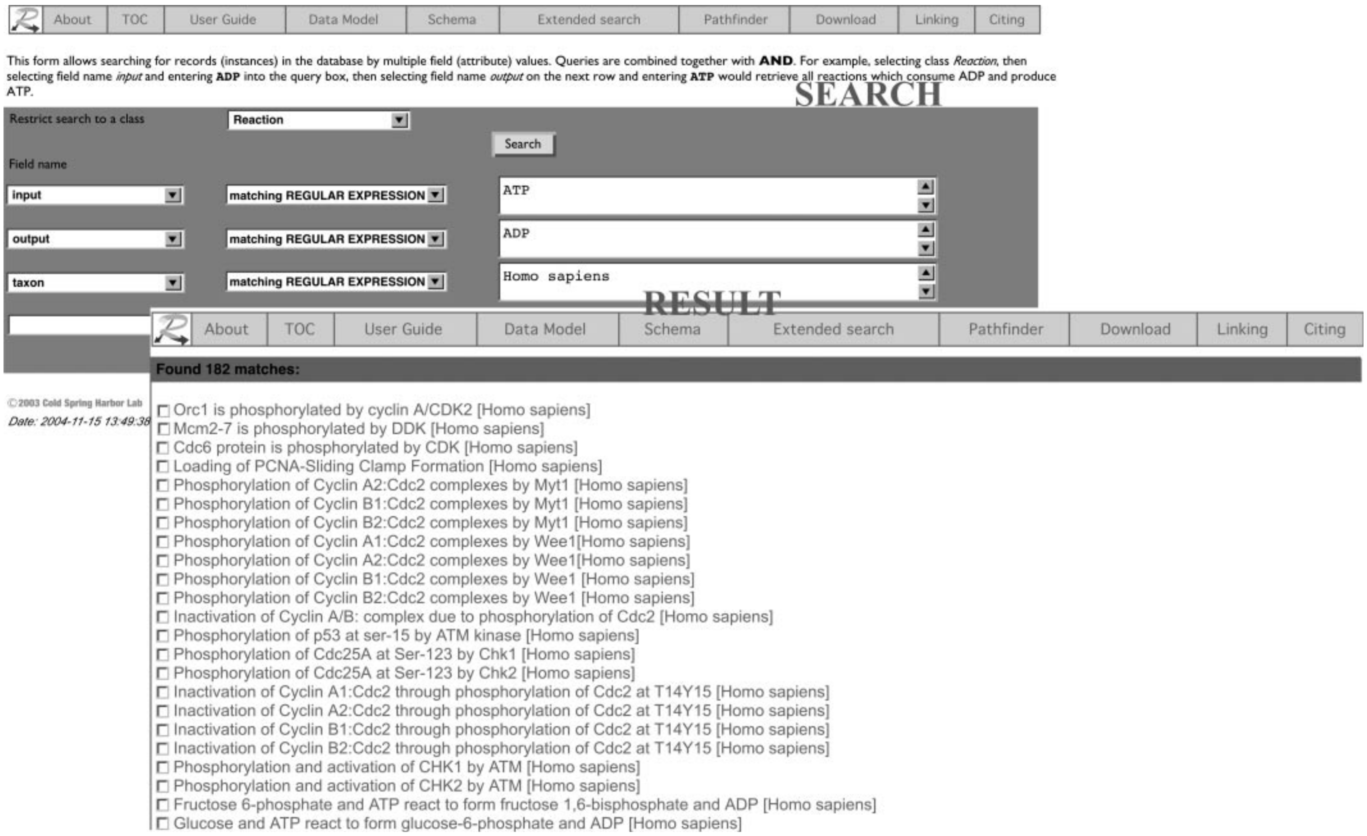
Reactome can also be used from the bottom up. Genomic databases provide comprehensive lists of genes and their protein products, but often with limited functional annotation. Reactome can be used to fill this gap for individual genes. A typical scenario occurs when a researcher encounters an unfamiliar UniProt protein, for example, a protein with the accession number Q14676 (Supplementary Material 6). Via agreement with UniProt, proteins whose functions have been annotated in Reactome have a link to the Reactome database. However, Q14676 is an entry in the TrEMBL database and only limited functional annotation is available for this protein. The link to Gene Ontology (GO) (3) only tells the user that this protein is intracellular, and is inferred by electronic annotation. On the other hand, the link to Reactome takes the researcher to a page that summarizes all the reactions and pathways in which the protein, a modification of the protein or a complex involving the protein, participates in (Supplementary Material 7), along with relevant literature citations and the module author's commentary (Supplementary Material 8). This immediately places the protein into its biological context in a richer and more useful manner than the simpler GO process annotation might provide. Similar to the UniProt links, Reactome also has reciprocal links to GO, Ensembl (4) and Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query. fcgi?db=Protein) databases.

Reactome features a full text search of the database as well as an advanced search. Advanced searches can be conducted across any Reactome data type and can be restricted by up to four properties. An example search is shown in Figure 2.

The Reactome 'skypainter' tool allows researchers to upload a list of gene or protein identifiers in order to colorize the reaction map in a number of ways. An example of the usage of this tool is shown in Figure 3, where the set of genes associated with human disease, the morbid map from Online Mendelian Inheritance in Man (OMIM) (5) (http://www.ncbi.nlm.nih.gov/omim/), has been overlaid onto the reaction map. The resulting graphic representation instantly shows that human disease genes are not arranged haphazardly, but instead cluster in certain key pathways. For example, ovarian cancer genes map onto the DNA damage checkpoint pathway, while those implicated in breast cancer can be found in the DNA double strand break repair pathway. Almost half the reactions in the reaction map are relevant to disease phenotypes.

Data model

The basic unit of Reactome is the *reaction*. A reaction is any event that converts inputs to outputs, where inputs and outputs are physical entities such as small molecules, proteins, lipids or nucleotides, or complexes of these. This definition of reaction is broad enough to encompass classical biochemical reactions, such as the phosphorylation of glucose to glucose-6-phosphate, as well as less conventional types of reactions, such as binding, dissociation, complex formation, translocation, polymerization and conformational changes. In addition to inputs and outputs, a reaction will include information on the species, sub-cellular location, and critically, the experimental



This form allows searching for records (instances) in the database by multiple field (attribute) values. Queries are combined together with **AND**. For example, selecting class *Reaction*, then selecting field name *input* and entering **ADP** into the query box, then selecting field name *output* on the next row and entering **ATP** would retrieve all reactions which consume ADP and produce ATP.

Restrict search to a class: Reaction ▼

Search

Field name		
input ▼	matching REGULAR EXPRESSION ▼	ATP
output ▼	matching REGULAR EXPRESSION ▼	ADP
taxon ▼	matching REGULAR EXPRESSION ▼	Homo sapiens

©2003 Cold Spring Harbor Lab
Date: 2004-11-15 13:49:38

Found 182 matches:

□ Orc1 is phosphorylated by cyclin A/CDK2 [Homo sapiens]
□ Mcm2-7 is phosphorylated by DDK [Homo sapiens]
□ Cdc6 protein is phosphorylated by CDK [Homo sapiens]
□ Loading of PCNA-Sliding Clamp Formation [Homo sapiens]
□ Phosphorylation of Cyclin A2:Cdc2 complexes by Myt1 [Homo sapiens]
□ Phosphorylation of Cyclin B1:Cdc2 complexes by Myt1 [Homo sapiens]
□ Phosphorylation of Cyclin B2:Cdc2 complexes by Myt1 [Homo sapiens]
□ Phosphorylation of Cyclin A1:Cdc2 complexes by Wee1[Homo sapiens]
□ Phosphorylation of Cyclin A2:Cdc2 complexes by Wee1[Homo sapiens]
□ Phosphorylation of Cyclin B1:Cdc2 complexes by Wee1 [Homo sapiens]
□ Phosphorylation of Cyclin B2:Cdc2 complexes by Wee1 [Homo sapiens]
□ Inactivation of Cyclin A/B: complex due to phosphorylation of Cdc2 [Homo sapiens]
□ Phosphorylation of p53 at ser-15 by ATM kinase [Homo sapiens]
□ Phosphorylation of Cdc25A at Ser-123 by Chk1 [Homo sapiens]
□ Phosphorylation of Cdc25A at Ser-123 by Chk2 [Homo sapiens]
□ Inactivation of Cyclin A1:Cdc2 through phosphorylation of Cdc2 at T14Y15 [Homo sapiens]
□ Inactivation of Cyclin A2:Cdc2 through phosphorylation of Cdc2 at T14Y15 [Homo sapiens]
□ Inactivation of Cyclin B1:Cdc2 through phosphorylation of Cdc2 at T14Y15 [Homo sapiens]
□ Inactivation of Cyclin B2:Cdc2 through phosphorylation of Cdc2 at T14Y15 [Homo sapiens]
□ Phosphorylation and activation of CHK1 by ATM [Homo sapiens]
□ Phosphorylation and activation of CHK2 by ATM [Homo sapiens]
□ Fructose 6-phosphate and ATP react to form fructose 1,6-bisphosphate and ADP [Homo sapiens]
□ Glucose and ATP react to form glucose-6-phosphate and ADP [Homo sapiens]

Figure 2. Screenshots of an example Advanced Search query are shown. The database was queried to find all reactions in *H.sapiens* that have ATP as input, ADP as output.

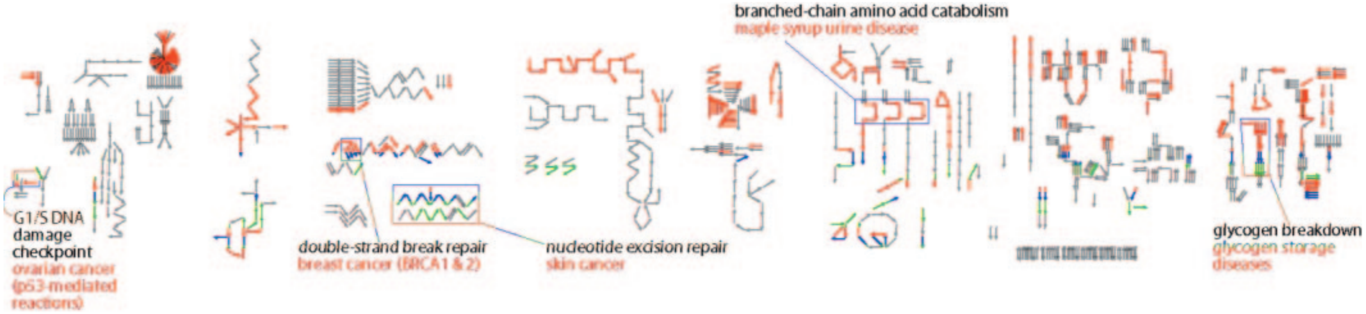

branched-chain amino acid catabolism
maple syrup urine disease
G1/S DNA damage checkpoint
ovarian cancer (p53-mediated reactions)
double-strand break repair
breast cancer (BRCA1 & 2)
nucleotide excision repair
skin cancer
glycogen breakdown
glycogen storage diseases

Figure 3. The OMIM Morbid Map of the Human Genome lists all genes whose mutant forms are causally associated with human disease. Each Reactome event in which, one or more such gene products are involved as input, catalyst or regulator is shown in red. Some examples of diseases that map to Reactome reactions are shown.

evidence for the reaction, typically taking the form of one or more literature citations. Other attributes of reactions include a catalyst activity, when appropriate, as well as information on their regulation. Reactions are then grouped into pathways that take into account their temporal relationships and interdependencies. Pathways in Reactome are useful groupings of reactions, and can contain sequential reactions, parallel reactions or reactions ordered in a cycle. Further, pathways can nest; pathways can have other pathways as their components, and can be sequential or parallel.

Many reactions are involved in the transformation of a physical entity from one state to another. For example, a carbohydrate transport reaction may move an extracellular sugar molecule into the cytosol. Reactome explicitly annotates such states by representing extracellular and cytosolic glucose as separate entries. Another example of the explicit annotation of states of a molecule is the p53 protein, which is represented by three distinct entities in Reactome: native p53, p53 phosphorylated at Ser15 and p53 phosphorylated at Ser20. This allows the distinct biological activities of each of these p53 states to be described unambiguously. These multiple states are derived from a single *Reference Entity*, which contains information on the polypeptide sequence of p53 as well as cross-references to the Uniprot, Entrez Gene and Ensembl databases.

The Reactome project is careful in using unambiguous, well-known identifiers whenever possible. In addition to links between reference entities and the protein and gene databases, Reactome links small molecules to ChEBI (http://www.ebi.ac.uk/chebi/), catalyst activities to the GO molecular function ontology, and sub-cellular locations to the GO cellular compartment ontology. These cross-references facilitate the integration of Reactome reactions and pathways with other bioinformatics Web resources. The data model also allows for statements about generic physical entities such as ‘any tRNA’ in order to avoid creating families of reactions that differ only by the particular species of tRNA that it operates on.

Data acquisition

Reactome is organized like an online journal. The editors, after consultation with the scientific advisory board, select a series of topics to annotate, and then invite bench biologists to author database ‘modules’ that are roughly the same scope and amount of work as a minireview. Our authors are typically at the faculty level, but run the gamut from postdoctoral fellows to tenured professors.

Authors create their modules using the Reactome Author Tool—a desktop application written in Java. This tool hides most of the complexities of the Reactome data model behind a graphical front end, whose major features are an interactive, user-friendly pathway editor and a task list pane that enforces consistency and completeness on the module. After the author has completed a module, it is handed over to a Reactome curator, who refines the annotation using a more sophisticated set of software applications. Reactome curators are full-time staff members who combine a broad knowledge of biology (most are PhD-level biologists) with a good understanding of bioinformatics and knowledge engineering. The curator’s job is to ensure that the module is complete and internally consistent.

After curation, the module appears on a private website for inspection by peer reviewers. Like the primary authors, peer reviewers are faculty-level biologists with expertise in the relevant field of biology. The authors and/or curators remedy any inconsistencies, omissions or errors discovered by the peer reviewers, and then the module is made available to the public. Topics are also on a schedule for a rolling review every two years, in order to keep the data current; the modules also get augmented with new data during such a review.

Non-human pathways in Reactome

Although Reactome is primarily concerned with curation and presentation of human processes, it is impossible to make assertions about human biology without reference to experimental work on non-human species. Non-human pathways appear in Reactome via two routes. The first route occurs when a non-human reaction is curated in order to provide indirect evidence for an implied equivalent human reaction. We prefer that our authors document reactions by using experiments performed on human systems (e.g. tissue culture). In many cases, however, a reaction is well described in yeast or frog, and only inferred in humans based on the finding of putative human orthologs for the proteins that participate in the reaction. In this case, the author creates a reaction for yeast or frog that is supported by direct experimental evidence, and then creates a ‘deduced’ reaction in human whose indirect evidence is the presence of the equivalent reaction in the model organism. By strictly separating direct and indirect evidence, we try to maintain clear chains of evidence and to leverage the power of comparative genomics without creating erroneous pathways in which the proteins of one species appear to interact with the proteins of another. Interactions between proteins from different species are only relevant while annotating inter-species relationships, such as the relationship between hosts and parasites or between symbionts, for example.

The second route by which non-human pathways appear in Reactome occurs just prior to each release, when we electronically project curated pathways from human onto each of the vertebrate species contained within the Ensembl Compara database (6). We build reactions in the non-human vertebrate where we can establish one-to-one mapping for putative orthologs based on mutual best hits between humans and a non-human vertebrate for the inputs or the catalysts of the reaction. This allows us to predict events in rat, mouse, fugu, zebra fish and chicken.

Current contents

Since the first announced release in January 2003, Reactome has grown to cover almost 10% of the human proteins in UniProt, and cites 894 literature references as supporting evidence for the reactions and pathways annotated in the database.

The number of proteins from the different species covered in Reactome, and the number of complexes, reactions and pathways they participate in, are listed in Table 1.

Currently, at version 10, we cover the following biological processes: cell cycle and its checkpoints, repair and replication of DNA, gene expression including the transcription by the three nuclear RNA polymerases, processing of the mRNA and its translation to protein, metabolism of sugars, ethanol, amino acids, nucleotides and lipids, the tricarboxylic acid cycle, and insulin receptor activation and recycling. We have completed a two-year rolling review of modules on DNA replication, and

Table 1. Reactome holdings

Species	No. of Proteins[a]	No. of Complexes	No. of Reactions	No. of Pathways
H.sapiens[b]	763	371	1247	296
Mus musculus	322	223	840	209
Rattus norvegicus	393	282	930	230
Danio rerio	258	209	680	200
Fugu rubripes	349	286	817	212
Gallus gallus	377	221	747	190
Total	2462	1592	5261	1337

[a] Proteins include entries referenced by UniProt and by Ensembl gene predictions.

[b] Data from humans is curated, whereas most of the data from the other organisms is from electronic event prediction.

we are currently in the process of similarly reviewing the modules on protein translation.

DISCUSSION AND FUTURE DIRECTIONS

The Reactome project is an attempt to capture all reactions and pathways thought to occur in humans. This is, however, only a small part of the story. In any cell or tissue type, only a small percentage of the whole genome is expressed, and therefore the full repertoire of the human reactome is never active simultaneously in any single cell or developmental stage. Knowledge of stage and tissue-specific expression patterns, along with kinetic data such as reaction rates and binding constants, are the key to creating quantitative models of physiology. Reactome does not attempt to capture stage or tissue-specific expression data, but defers this task to other databases that are capturing the results of high-throughput expression studies. We see the role of Reactome as providing a framework of possible reactions which, when combined with expression and enzyme kinetic data, provides the infrastructure for quantitative models. In order to facilitate this type of data integration, we are working to make Reactome data available in a variety of standard formats, including BioPAX, SBML and PSI-MI. This will also enable data exchange with other pathway databases, such as the Cycs (7), KEGG (8) and amaze (9), and molecular interaction databases, such as BIND (10) and HPRD (11).

The next data release, version 11, will cover apoptosis, including the death receptor signaling pathways, and the Bcl2 pathways, as well as pathways involved in hemostasis. Other topics currently under development include several signaling pathways, mitosis, visual phototransduction and hematopoeisis.

In summary, Reactome provides high-quality curated summaries of fundamental biological processes in humans in a form that is as useful to students working on a single protein as it is to bioinformaticists striving to make sense of large-scale datasets. Reactome provides biologist-friendly visualization of biological pathways data, and is an open-source project.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to our external authors and peer-reviewers who help us to maintain a high quality of data in Reactome. The development of Reactome is supported by grants R01 HG002639 from the NHGRI at the US NIH and LSHG-CT-2003-503269 from EU (6th Framework Programme) and a subcontract from the NIH-funded Cell Migration Consortium.

REFERENCES

1. Joshi-Tope,G., Vastrik,I., Gopinathrao,G., Matthews,L., Schmidt,E., Gillespie,M., D'Eustachio,P., Jassal,B., Lewis,S., Wu,G. *et al.* (2003) The Genome Knowledgebase: a resource for biologists and bioinformaticists. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 237–243.
2. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
3. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
4. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl. *Nucleic Acids Res.*, **32**, D468–D470.
5. McKusick,V.A. (1998) *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, 12th edn. Johns Hopkins University Press, Baltimore, MD.
6. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
7. Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.
8. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
9. Lemer,C., Antezana,E., Couche,F., Fays,F., Santolaria,X., Janky,R., Deville,Y., Richelle,J. and Wodak,S.J. (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.*, **32**, D443–D448.
10. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
11. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, **32**, D497–D501.